베이지안망을 이용한 유전자 발현 데이터의 분석

황규백 장병탁 김영택

서 울 대 학 교 컴 퓨 터 공 학 부

kbhwang@scai.snu.ac.kr {btzhang, ytkim}@cse.snu.ac.kr

Gene Expression Data Analysis Using Bayesian Networks

Kyu-Baek Hwang Byoung-Tak Zhang Yung Taek Kim

School of Computer Science and Engineering, Seoul National University

요 약

최근 DNA 칩 또는 마이크로어레이 기술의 발전으로 인해 한 세포 내의 수천 개의 유전자의 발현 정도를 동시에 측정할 수 있게 되었다. 이러한 마이크로어레이 데이터를 분석해서 암의 경과나 세포의 주기적 변화 등에 영향을 미치는 유전자들을 알아낼 수 있다. 본 논문에서는 베이지안망을 이용해서 마이크로어레이 데이터를 분석, 백혈병의 경과를 예측한다. 베이지안망은 다수의 변수들간의 확률적 관계를 표현하는 그래프 모델로 각 유전자들간의 확률적 관계를 사람이 알아보기 쉬운 형태로 학습할 수 있다는 장점이 있다. 마이크로어레이 데이터에 대해서 학습된 베이지안망은 백혈병 경과 예측에 대해서 기존의 방법보다 뛰어난 성능을 보였다.

1. 서론

생물의 세포(cell)는 종에 따라 수천, 수만 개의 유전자(gene)를 가지고 있다. 이러한 유전자들은 주변 상황에 따라 적절한 단백질(protein)을 합성하는데 필요한 정보들을 담고 있다. 유전자 발현(gene expression)은 유전자가 지니는 정보가 사용되기 위해 일어나는 일련의 과정이다. DNA hybridization은 세포 내의 수천 개의 유전자들의 발현 정도를 동시에 측정하는 기술로, 이렇게 측정된 유전자 발현 데이터를 마이크로어레이(microarray) 데이터라 한다. 이러한 마이크로어레이 데이터를 분석해서 유용한 정보를 얻으려는 노력이 있어 왔다. 예를 들어 [7], [3]은 효모 세포의 주기적 변화에 따라 달라지는 유전자들의 발현 정도를 분석해서 세포의 주기에 영향을 주는 유전자들간의 관계를 밝혀냈다. 또한 [6]은 마이크로어레이 데이터를 분석해서 백혈병(leukemia)의 경과에 영향을 주는 유전자들을 밝혀내고 이를 이용해서 환자의 세포에서 추출된 유전자 발현 데이터로부터 백혈병의 경과를 예측했다.

마이크로어레이 데이터의 분석에는 통계적 기법과 기계학습의 기법들이 이용되어 왔다. 본 논문에서는 베이지안망(Bayesian network)으로 [5]의 백혈병 마이크로어레이 데이터를 분석한다.

급성 백혈병(acute leukemia)은 크게 AML(acute myeloid leukemia)과 ALL(acute lymphoblastic leukemia)로 구분된다. AML과 ALL은 각기 병의 경과가 다르므로 이의 구분은 병의 진단과 치료에 필수적이다. 그러나 조직 검사만으로는 이 구분이 쉽지 않으며 유전자 발현 데이터의 분석이 유용하다는 견해가 있어 왔다. [5]의 데이터는 38명의 백혈병 환자들의 7129개의 유전자의 발현 정도를 측정한 데이터이다. 이 논문에서는 백혈병을 구분할 수 있는 베이지안망 분류기(Bayesian network classifier)를 마이크로어레이 데이터에 대해서 학습한다. 베이지안망 분류기는 백혈병의 분류뿐 아니라 백혈병의 구분에 영향을 주는 유전자들간의 확률적 관계도 학습할 수 있는 장점이 있다.

2. 베이지안망 (Bayesian networks)

베이지안망은 변수에 해당하는 노드와 그 노드(변수)들간의 확률적 관계를 나타내는 방향성 간선들로 구성된 DAG(directed acyclic graph) 구조를 가지며 변수들간의 결합확률분포(joint probability distribution)를 효율적으로 표현할 수 있는 그래프 모델이다. 변수 집합 $\mathbf{X} = \{X_1, \ldots, X_n\}$에 대한 베이지안망은 다음의 2가지 부분으로 구성된다.

(1) $\mathbf{X}$의 변수들간의 조건부독립성(conditional independence assertion)을 표현하고 있는 망 구조 S

(2) 각 변수들의 지역확률분포(local probability distribution) 집합 P

망 구조 S는 DAG 형태이며 S의 각 노드는 $\mathbf{X}$의 변수들과 일대일대응이 된다. X_i는 변수와 그 변수에 해당하는 노드를 동시에 가리킨다. $\mathbf{Pa}_i$는 그래프 S에서 X_i의 부모노드(변수)의 집합을 나타낸다. S에서 간선으로 연결되지 않은 노드들은 서로 조건부독립관계에 있다. 망 구조가 나타내는 조건부독립성에 의하면 주어진 구조 S에서 $\mathbf{X}$의 결합확률분포는 다음과 같이 표현된다.

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid \mathbf{pa}_i)$$

지역확률분포 P는 위 수식의 Π 안의 각 항에 대응된다. 베이지안망 (S, P)가 주어지면 원하는 확률을 추론할 수 있다.

3. 베이지안망을 이용한 백혈병 마이크로어레이 데이터의 분석

베이지안망을 이용한 유전자 발현 데이터 분석의 목적은 각 유전자들 간의 확률적 관계를 학습하는 것이다. 따라서 하나의 유전자는 베이지안망의 하나의 노드에 해당된다. 또한 백혈병의 구분을 위한 leukemia class라는 노드가 추가된다. leukemia class 변수는 백혈병의 종류인

AML, ALL의 값을 가진다. 이렇게 구성된 베이지안망을 이용해 각 유전자의 발현값이 주어진 경우의 leukemia class 변수의 확률을 추론할 수 있다. 이를 이용해서 유전자의 발현 정보를 가지고 백혈병의 종류를 구분할 수 있다.

3.1 유전자 발현값의 이산화 (Discretization of Gene Expression Levels)

베이지안망의 적용을 위해서 우선 각 노드의 지역확률분포 모델을 설정해야 한다. 베이지안망의 지역확률분포 모델로는 다항분포와 선형 가우시안 분포가 많이 이용된다. 본 논문에서는 다항분포를 이용하였다. 유전자 발현값은 실수값이기 때문에 다항분포를 이용하기 위해서는 이를 이산화(discretization)해야 한다. 이산화 방법에는 여러가지가 있으며[2] 이 논문에서는 Equal Width Interval Binning[2]과 각 유전자 발현값의 평균을 기준으로 0과 1로 나누는 방법을 사용하였다. Equal Width Binning은 단순하게 데이터의 최대값과 최소값을 이용해서 그 사이의 값들을 동일한 크기의 구간으로 구분하는 방법이다. 평균을 기준으로 0과 1로 나누는 방법은 데이터의 유전자 발현값들이 정규화된 형태[5]이기 때문에 의미가 있다. 즉, 평균 이상의 발현값을 가지는 경우를 1로 그렇지 않은 경우를 0으로 설정하는 것이다.

3.2 백혈병 구분에 관련된 유전자의 선정 (Selection of Genes Related to the Classification of Leukemia)

[5]의 데이터는 7129개의 유전자의 발현 정도를 측정한 38개의 학습 예제(training example)로 구성되어 있다. 유전자의 개수가 학습 예제에 비해 절대적으로 많기 때문에 이를 모두 이용한 베이지안망의 학습은 어렵다. 따라서 백혈병 구분과 관련이 있는 적절한 수의 유전자를 선정해야 한다. [6]에서는 *P*-metric이라는 기준을 이용해서 유전자를 선정했다. 본 논문에서는 유전자와 leukemia class 변수의 상호 정보량(mutual information)을 이용해서 유전자들을 선정했다. 상호 정보량을 이용해서 유전자를 선정하는 방법은 두 가지이다. 하나는 상호 정보량의 크기 순으로 일정 개수의 유전자를 선정하는 것이다. 다른 방법은 AML의 경우에 발현하는 유전자와 ALL의 경우에 발현하는 유전자를 구분하고 각 경우에서 상호 정보량의 순으로 유전자를 선정하는 것이다.

유전자의 선정 방법이 정해졌으면 베이지안망에 포함시킬 유전자의 개수를 정해야 한다. 이 개수는 학습 예제의 개수와 관련이 있다. 베이지안망이 학습하는 확률이 어느 정도의 신뢰도를 가지려면 학습 예제의 수가 일정량 이상이 되어야 한다. 이 논문에서 이용한 데이터의 학습 예제의 개수는 38개로 베이지안망의 노드의 개수가 6개 이상이 되는 경우에는 베이지안망이 학습한 확률의 신뢰도가 너무 낮게 된다. 따라서 노드의 개수는 5개로 제한했다. 5개의 노드 중 하나는 leukemia class 노드이므로 선정되는 유전자의 개수는 4개가 된다.

3.3 베이지안망의 학습

학습에 사용할 유전자가 선정되었으면 유전자 노드와 leukemia class 노드를 가지는 베이지안망의 구조를 학습해야 한다. 베이지안망 구조가 학습 데이터에 적합한 정도를 베이지안망의 점수라 한다. 베이지안망의 구조 학습은 점수가 높은 망 구조를 찾는 과정으로 이루어진다. 베이지안망의 점수로는 BDe (Bayesian Dirichlet metric) 점수[4]를 이용했다. 이 점수는 다음의 식으로 표현된다.

$$p(D,B) = p(B)\cdot\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})}$$

위 식에서 D는 학습 데이터, B는 베이지안망의 구조이다. n은 학습 예제의 개수, q_i는 노드 i의 부모 노드 집합이 가질 수 있는 상태의 개수이며 r_i는 노드 i의 상태의 개수이다. N_{ijk}는 학습 데이터에서 노드 i가 부모노드의 j번째 상태 하에서 k번째 상태를 가지는 횟수이다. α_{ijk}는 노드 i의 Dirichlet prior로 실험에서는 1.0의 값[4]이 이용되었다. α_{ij}와 N_{ij}는 다음과 같이 계산된다.

$$\alpha_{ij} \equiv \sum_{k=1}^{r_i}\alpha_{ijk} \qquad N_{ij} \equiv \sum_{k=1}^{r_i} N_{ijk}$$

위의 점수에서 $p(B)$는 망 구조에 대한 사전 확률(prior probability)이다. 망의 구조에 대한 사전 지식은 가지고 있지 않다고 가정하고 모든 망 구조의 사전 확률은 동일하게 주었다. $\Gamma(\cdot)$ 함수의 값이 매우 커질 수 있기 때문에 실험에서는 위의 점수에 log 함수를 취한 값을 사용하였다.

n개의 노드를 가지는 베이지안망 구조의 탐색 공간 크기는 다음과 같다.

$$n! \times 2^{\frac{n(n-1)}{2}}$$

탐색 공간이 매우 크며 일반적인 경우, 가장 좋은 베이지안망 구조의 탐색은 NP-hard 문제이다[1]. 따라서 주로 greedy search algorithm이 이용된다. 본 논문에서는 베이지안망 노드의 개수를 5개로 제한했으므로 모든 공간을 탐색하는 방법을 사용했다.

4. 실험 결과

4.1 선정된 유전자 (Selected Genes)

3.1절에서 설명한 두 가지의 이산화(discretization) 방법과 3.2절에서 설명한 두 가지의 유전자 선정 방법에 따라 선정된 유전자들은 표 1과 같다. 표 1에서 유전자 옆 괄호 속의 숫자는 leukemia class 변수와의 상호 정보량값(mutual information value)이다.

표 1. 각 방법에 따라 선정된 유전자들

	Equal Width Binning	평균기준구분
단순한 상호 정보량 순	유전자 집합 1 FAH (0.737361) LTC4S (0.497781) Liver mRNA for IGIF (0.471081) HoxA9 (0.442018)	유전자 집합 2 Zyxin (0.704252) ADM (0.591874) LTC4S (0.501848) NADPH (0.497781)
AML과 ALL의 구분	유전자 집합 3 FAH (0.737361) LTC4S (0.497781) Platelet IB (0.311731) N-Methyl (0.265755)	유전자 집합 4 Zyxin (0.704252) ADM (0.591874) C-myb (0.37707) MB-1 (0.316313)

4.2 베이지안망의 성능 평가

표 1의 유전자 집합들에 대해서 4개의 베이지안망을 학습한 뒤 각각의 성능을 평가해 보았다. 베이지안망의

백혈병 구분에 대한 성능 평가는 학습 데이터와 테스트 데이터에 대해서 행했다. 테스트 데이터는 학습 데이터와는 별도로 34개의 예제로 구성되어 있다. leukemia class 변수의 확률이 0.5가 되는 경우는 에러로 간주했으며 각 베이지안망의 성능은 표 2에 나와 있다.

표 2. 베이지안망의 성능

	학습 에러	테스트 에러
유전자 집합 1	**0/38**	10/34
유전자 집합 2	**0/38**	9/34
유전자 집합 3	1/38	4/34
유전자 집합 4	1/38	**2/34**

표 2를 보면 유전자 집합 1, 2의 경우에는 학습 에러는 0이지만 테스트 데이터에 대한 에러는 상당히 큼을 알 수 있다. 유전자 집합 1, 2는 단순히 상호 정보량의 크기에 따라 선정된 유전자들이다. 반면에 유전자 집합 3, 4는 AML의 경우에 발현하는 유전자들과 ALL의 경우에 발현하는 유전자들이 각각 상호정보량에 따라 선정된 집합이다. 이 결과에 따르면 백혈병의 구분에 필요한 유전자를 찾기 위해서는 백혈병의 종류에 따라 발현하는 유전자들을 고르게 고려해야 한다는 사실을 알 수 있다. 또한 유전자 집합 4의 테스트 데이터에 대한 에러가 가장 적은 것을 보면 이산화 방법으로는 평균을 기준으로 0과 1로 나누는 것이 더 적절함을 알 수 있다. 참고로 [6]에서는 6817개의 유전자 중 50개를 사용했으며 동일한 데이터에 대해서 학습 에러는 2/38, 테스트 에러는 5/34였다. 그림 1은 테스트 데이터에 대한 성능이 가장 좋은 유전자 집합 4에 대한 베이지안망의 구조이다.

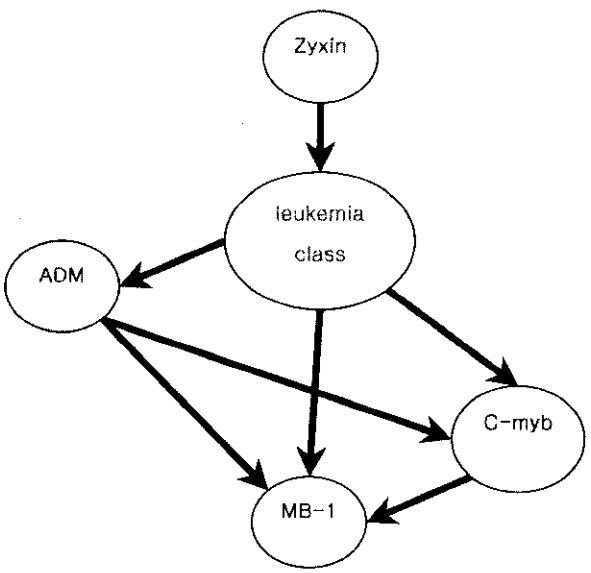

그림 1. 유전자 집합 4로 구성된 베이지안망의 구조

5. 결론

본 논문에서는 상호 정보량(mutual information)을 이용해서 백혈병의 구분에 영향을 주는 유전자를 찾아내고 이를 이용하여 백혈병의 종류 AML, ALL을 구분할 수 있는 베이지안망 분류기(Bayesian network classifier)를 학습했다. 특히 유전자 집합 3, 4로 구성된 베이지안망 분류기는 기존의 방법[6]보다 우수한 성능을 보였다. 또한 수천 개의 유전자 중 불과 4개의 유전자 발현 정보로 백혈병의 종류를 구분할 수 있다는 사실을 보였다.

베이지안망의 장점 중 하나는 학습 결과를 사람이 이해하기 쉽다는 점이다. 그림 1의 베이지안망은 각 유전자와 백혈병 종류간의 확률적 관계를 나타내고 있으며 그래프의 간선은 각 유전자간의 인과관계의 가능성을 나타내고 있다[3]. 이는 베이지안망이 유전자망(gene network)의 구성에 적용 가능하다는 사실을 나타낸다. 베이지안망의 구조에만 관심이 있는 경우는 적은 수의 데이터로도 더 많은 수의 유전자에 대한 베이지안망을 구성할 수 있으며 여기에는 bootstrap과 같은 통계적 기법이 적용될 수 있다[3].

감사의 글

이 논문은 교육부 BK21 사업에 의하여 지원되었음.

참고 문헌

[1] Chickering, D. M., Learning Bayesian networks is NP-complete, *Lecture Notes in Statistics*, 1995.

[2] Dougherty, J., Kohavi, R., and Sahami, M., Supervised and Unsupervised Discretization of Continuous Features, In *Proceedings of ICML'95*, pp. 194-202, 1995.

[3] Friedman, N., Linial, M., Nachman, I., and Pe'er, D., Using Bayesian Networks to Analyze Expression Data, In *Proceedings of RECOMB'00*, pp. 127-135, 2000.

[4] Heckerman, D., Geiger, D., and Chickering, D. M., Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, vol. 20, pp. 197-244, 1995.

[5] http://waldo.wi.mit.edu/MPR

[6] Slonim, D. K. et al., Class Prediction and Discovery Using Gene Expression Data, In *Proceedings of RECOMB'00*, pp. 263-272, 2000.

[7] Spellman, P. T. et al., Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization, *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.